Internet Archive: Universal Access. Open APIs.

linux.conf.au 2016
Open Knowledge Mini-Conf

VM Brasseur

lca2016@vmbrasseur.com
@vmbrasseur
IRC: vmbrasseur
http://vmbrasseur.com

Hewlett Packard Enterprise

Alexis Rossi

Director of Media & Access
Internet Archive
alexis@archive.org
@alexisrossi
http://archive.org

INTERNET ARCHIVE

Vicky Brasseur

lca2016@vmbrasseur.com
@vmbrasseur
IRC: vmbrasseur
http://vmbrasseur.com

Hewlett Packard Enterprise

Alexis Rossi

Director of Media & Access
Internet Archive
alexis@archive.org
@alexisrossi
http://archive.org

INTERNET ARCHIVE

Rex Features

Some Handy Links

- https://github.com/internetarchive
- https://archive.org/details/lca2016-ia-apis
- https://www.zotero.org/groups/internet_archive_-_open_apis_and_examples/items

A Tour of Internet Archive

Funston
CHRISTIAN SCIENCE HEALS

Internet Archive

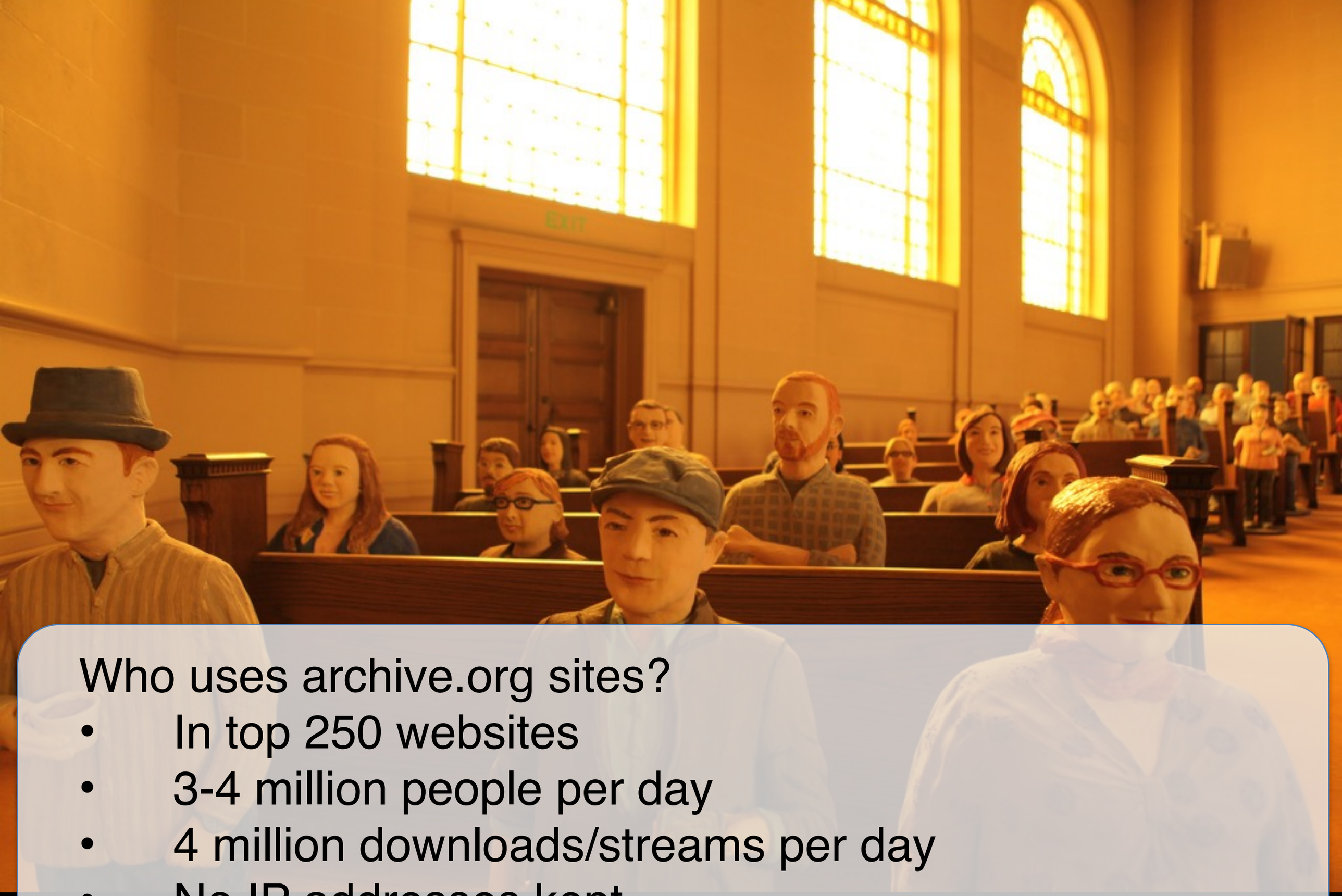
- 501(c)(3) non-profit, founded in 1996
- Registered library in California, USA
- Partnerships with hundreds of libraries, museums & universities
- 24 petabytes of unique data

Photo: Jason Scott

Where does the money come from?

- 40% digitization projects
- 20% web archiving projects
- 40% foundations & donations

Photo: Jason Scott



Who uses archive.org sites?

- In top 250 websites
- 3-4 million people per day
- 4 million downloads/streams per day
- No IP addresses kept

Photo: Jason Scott

Archived web sites

- 400+ billion web captures, 1996 to present
- Wayback Machine updated within hours
- 700,000 people per day
- 80TB crawl open for bulk download

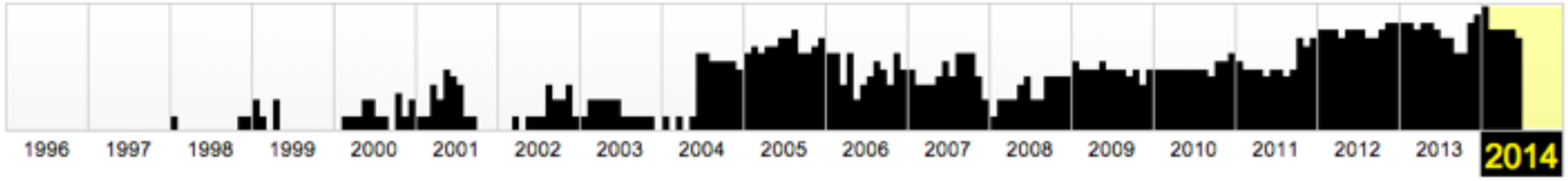
INTERNET ARCHIVE
WayBackMachine

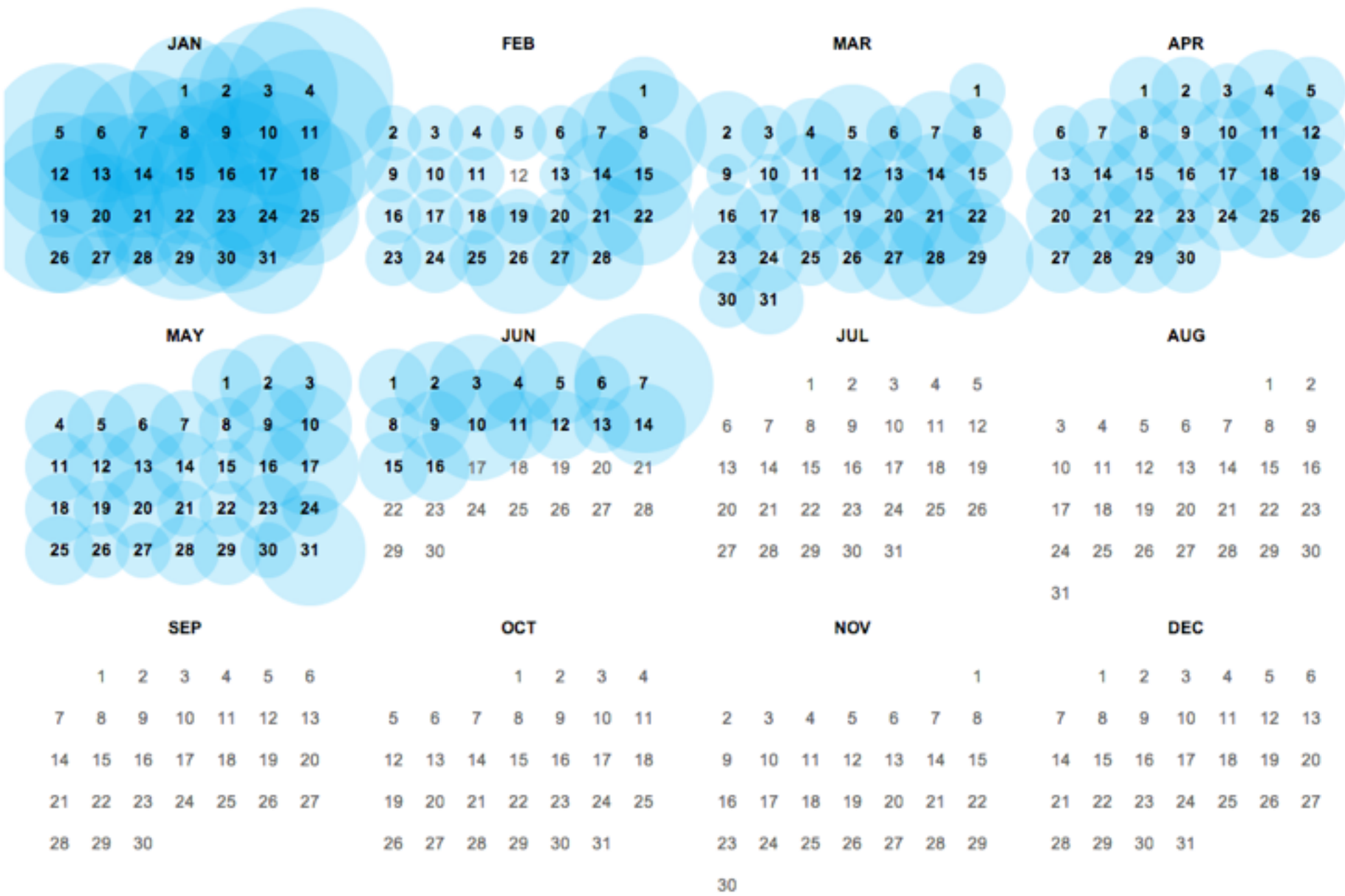
http://www.gnu.org/ BROWSE HISTORY

http://www.gnu.org/

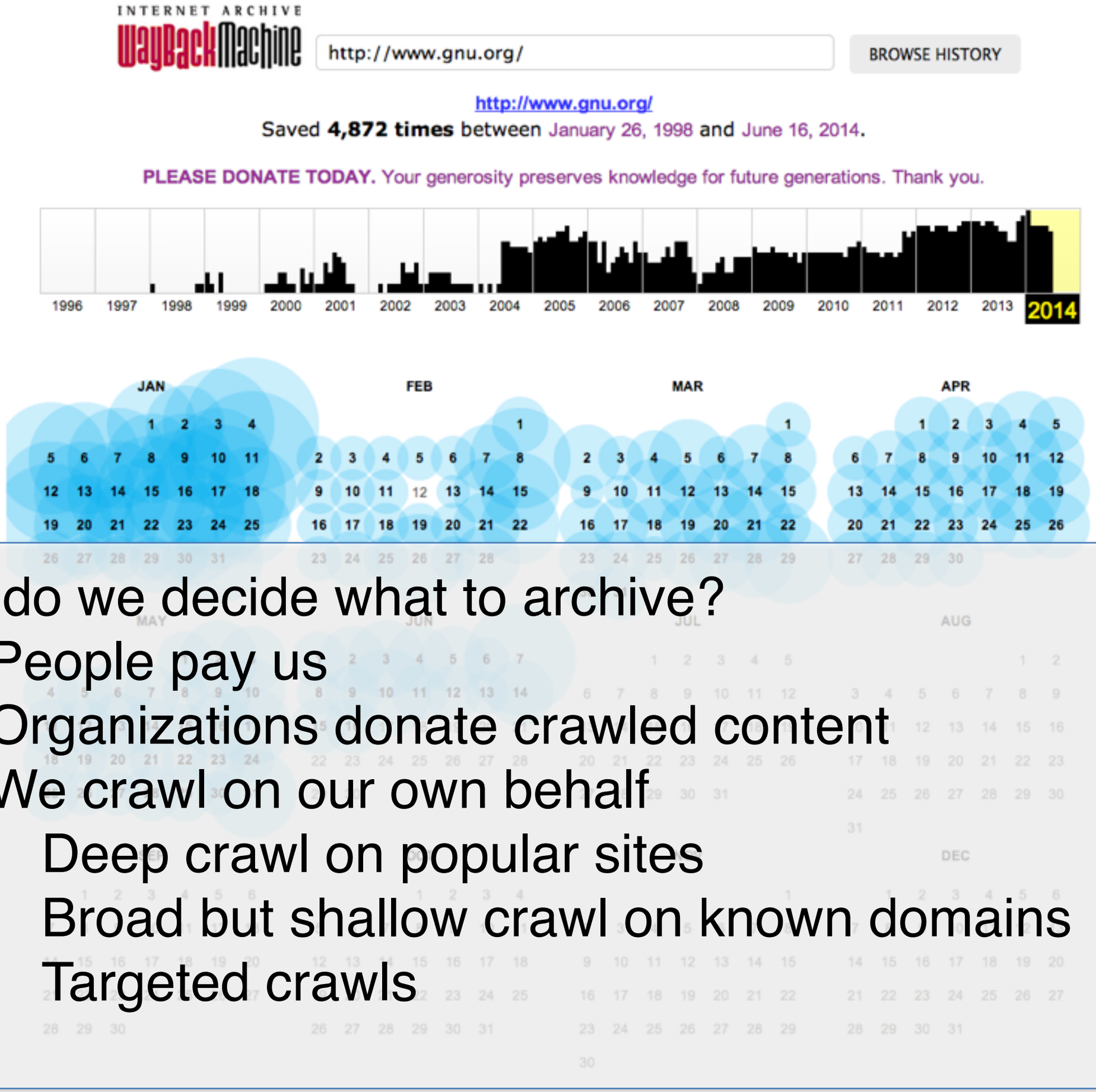
Saved **4,872 times** between January 26, 1998 and June 16, 2014.

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.





Note



How do we decide what to archive?

- People pay us
- Organizations donate crawled content
- We crawl on our own behalf
  - Deep crawl on popular sites
  - Broad but shallow crawl on known domains
  - Targeted crawls

VMBrasseur

INTERNET ARCHIVE
WayBackMachine

http://

BROWSE HISTORY

432 billion web pages saved over time. DONATE

Tools

Wayback Machine Availability API
Build your own tools.

WordPress Broken Link Checker
Banish broken links from your blog.

404 Handler for Webmasters
Help users get where they were going.

Subscription Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.

Save Page Now

http://

SAVE PAGE

Capture a web page as it appears now for use as a trusted citation in the future.

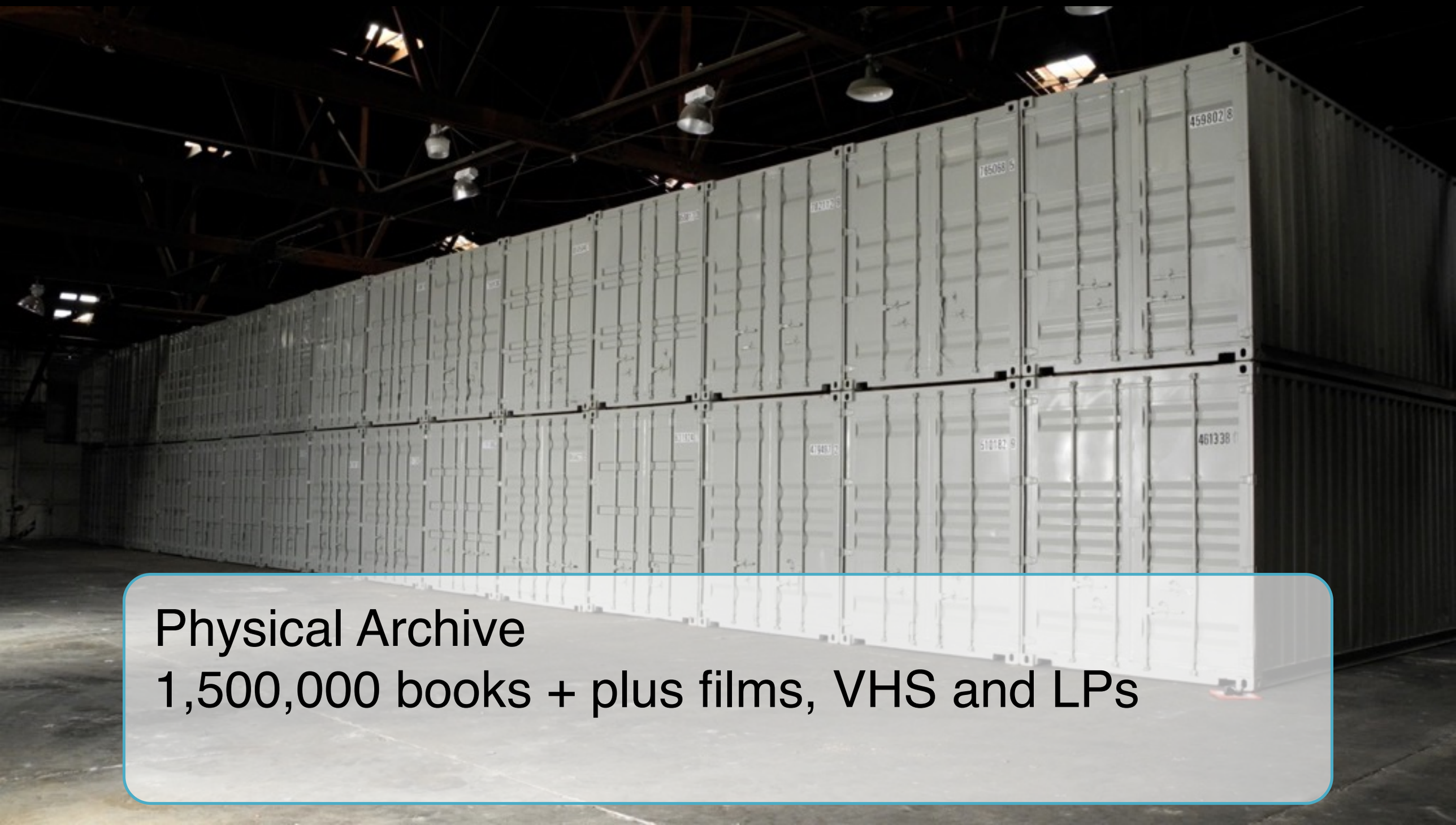
Only available for sites that allow crawlers.

MEKEL

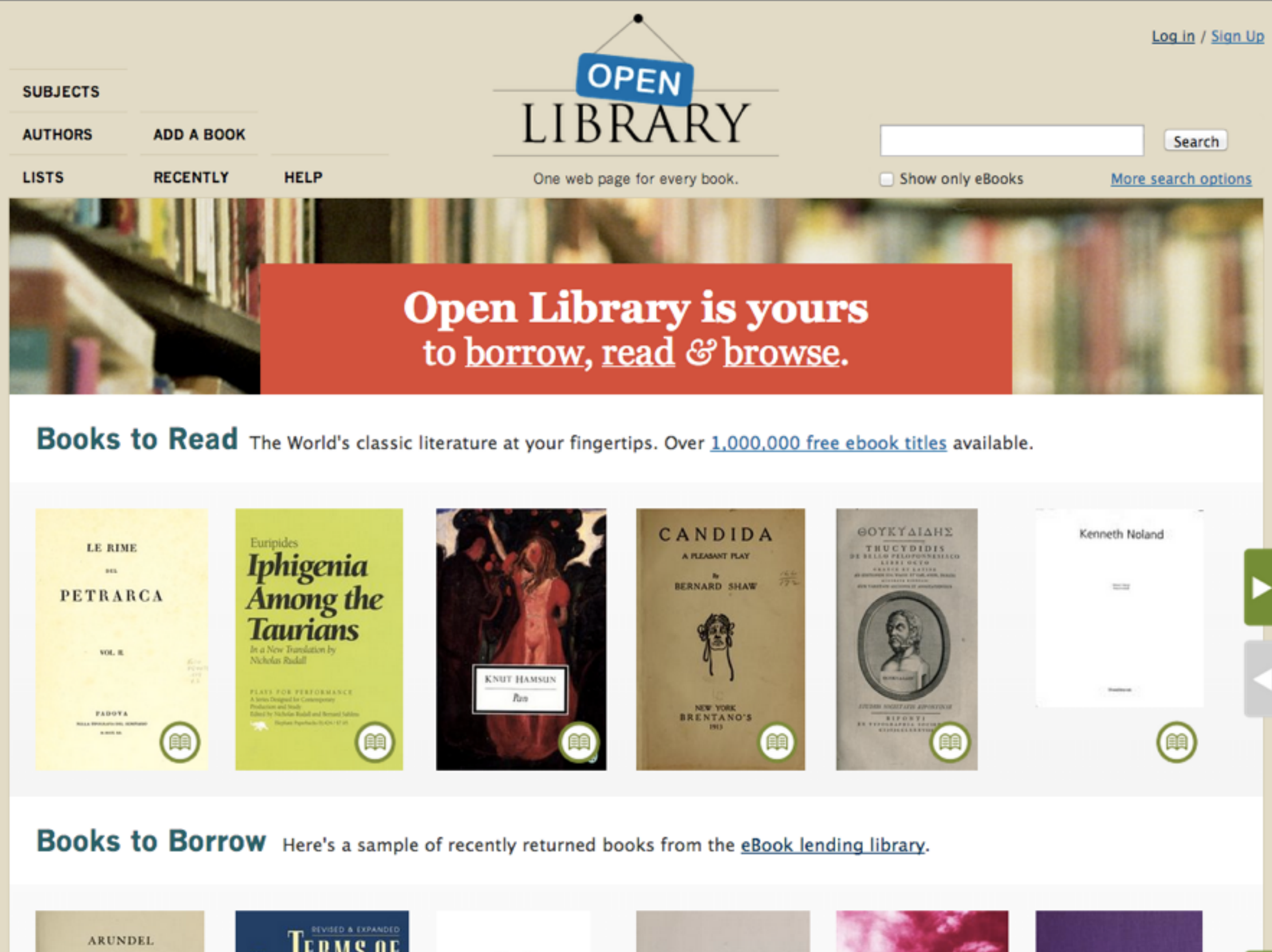
- 30 scanning centers in 8 countries
- 1,000 books per day

November 2013

459802 8
461338



Physical Archive
1,500,000 books + plus films, VHS and LPs



Borrow books at openlibrary.org

Video

- 2,000,000 items
- Feature films, documentaries, commercials, propaganda, stock footage, cartoons
- Archive 65 channels of TV 24 hours per day

INTERNET ARCHIVE Web Video Texts Audio Software TVNews OpenLibrary About Account Upload

TV NEWS ARCHIVE ukraine Search inside show Guide Me

Al Jazeera America | News | In-depth coverage of the top international stories.

May 5, 2014
8:00 - 9:01pm EDT

Borrow a DVD of this show

8:02pm

GRID SCAN

8:00pm

that's why the show will be talking about the indian elections and economy in the days and weeks to come. that is our show for today. i'm ali velshi. thanks for joining us. good evening, everyone, this is al jazeera america. i'm john siegenthaler in new york. nigerian nightmare a new

8:01pm

crisis. and how swapping blood of the young with the old may reverse the aging crisis. and cinco de mayo the true story behind the unusual holiday. ♪ we begin in nigeria where the anguish has intensified for the families of 276 girls kidnaps three weeks ago. richelle carey is here with more. >>

about 300 young women ages 16 to 18 were told they would be taken somewhere safe by men dressed in police uniforms. but the men were actually members of boko haram. we know little about the girls. only some of their names have been released.

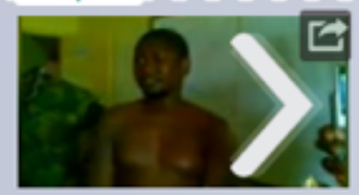
8:03pm

back our daughters. a campaign that began on twitter and quickly spread. some protesters in nigeria say they are being targeted by the government. yvonne ndege says accusations could stir up more anger. >> reporter: these allegations could lead more people to the streets. >> reporter:

8:04pm

everything possible to support nigeria and return the young pem to their homes. >> how they disappeared without anybody seeing anything has been the question all along. nigeria has been battling boko haram for years now. the group's name means western education is sinful. randall pinkston

8:05pm

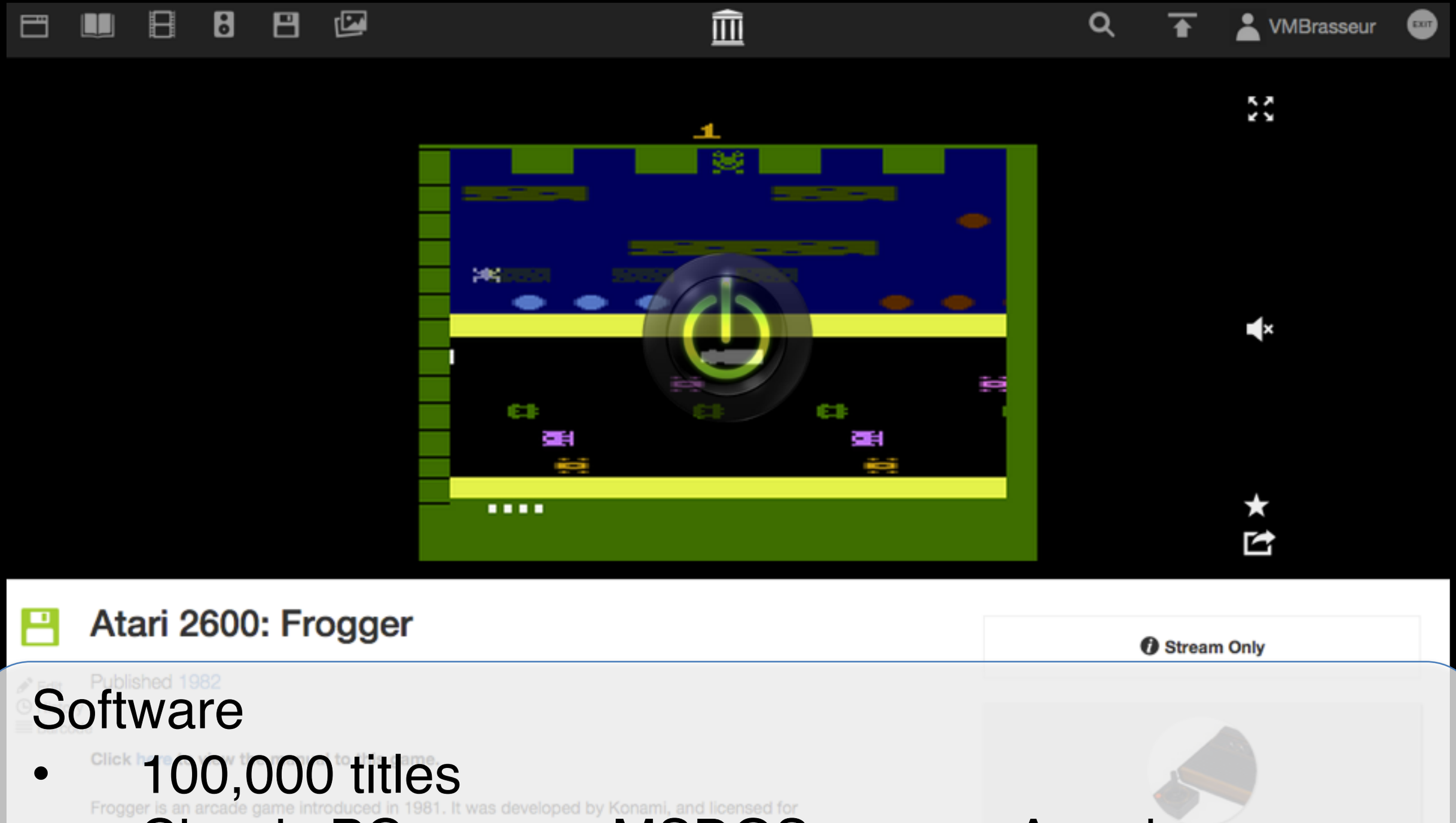


group. >> boko haram is a frig movement. it is a terrorist group which has assassinated muslim learned men, terrorized communities, and it is a small group relatively speaking but casts large shadow. >> reporter: its latest move kidnapping hundreds of schoolgirls and

U.S. TV News Archive: archive.org/tv

Audio

- 2,500,000 items
- 150,000 live concerts from 6,000 bands
- 60,000 netlabel releases from 2,000 labels
- Mirrors of Jamendo, IUMA
- Audio books, radio shows, podcasts



Software

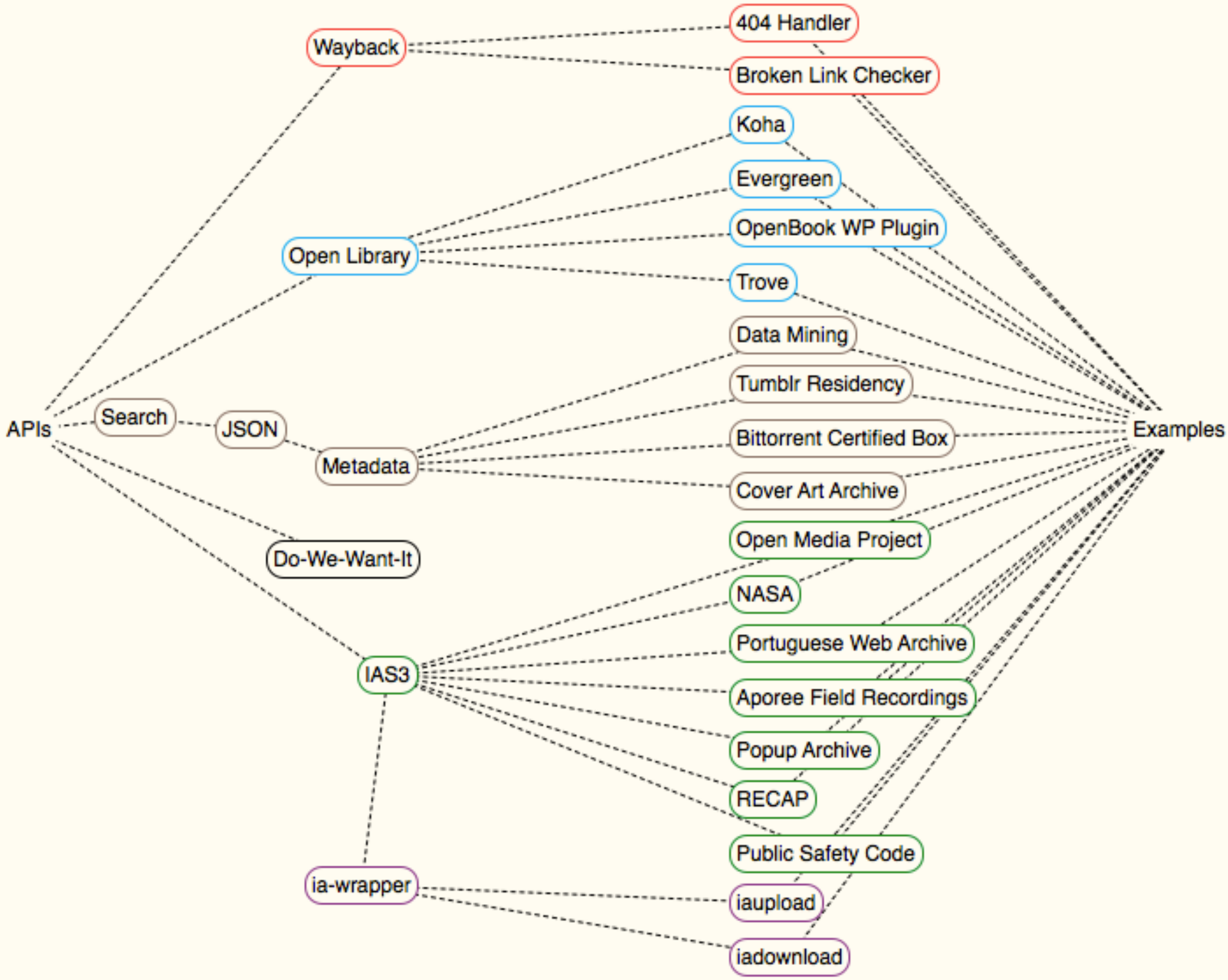
- 100,000 titles
- Classic PC games, MSDOS games, Arcade, Consoles
- Emulation in browser

Talk to Jason Scott, jscott@archive.org

Why?

- We are good at storing and serving digital media and preserving it
- We care about the same things: knowledge, keeping information open, privacy
- We fight for what we care about
- We’re not slick, but we are friendly!

Internet Archive APIs


APIs
Wayback
Open Library
Search
JSON
Metadata
Do-We-Want-It
IAS3
ia-wrapper
404 Handler
Broken Link Checker
Koha
Evergreen
OpenBook WP Plugin
Trove
Data Mining
Tumblr Residency
Bittorrent Certified Box
Cover Art Archive
Open Media Project
NASA
Portuguese Web Archive
Aporee Field Recordings
Popup Archive
RECAP
Public Safety Code
iaupload
iadownload
Examples

Wayback Machine API

http://archive.org/help/wayback_api.php

Wayback Machine API

- Is a URL archived?
- If so, is it available in the Wayback Machine?

Wayback Machine API

```
http://archive.org/wayback/available?
PARAMETERS
```

Wayback Machine API

Parameters:

url=sub.domain.tld
timestamp=20030831060429 (YYMMDDhhmmss; optional)
callback=function_name (optional)

National Aeronautics and Space Administration

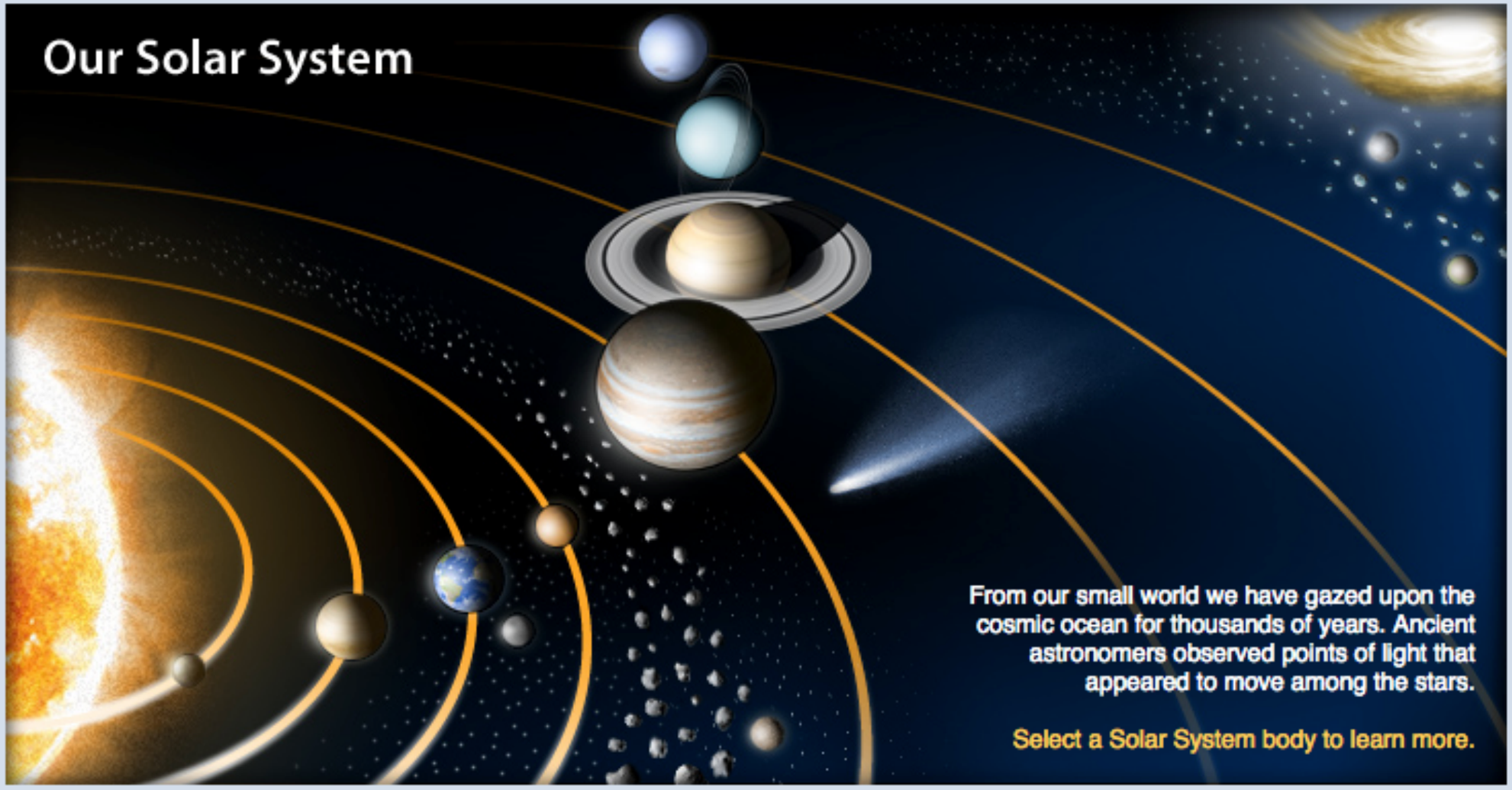
Solar System Exploration

Planets

- Home
- News & Events
- Planets
  - What is a Planet?
  - Our Solar System
  - Sun
  - Mercury
  - Venus
  - Earth
  - Earth's Moon
  - Mars
  - Asteroids
  - Meteors & Meteorites
  - Jupiter
  - Saturn
  - Uranus
  - Neptune
  - Dwarf Planets
  - Comets
  - Kuiper Belt & Oort Cloud
  - Beyond Our Solar System

SSE Home > Planets

Planets



A solar system refers to a star and all the objects that travel in orbit around it. Our solar system consists of the sun - our star - eight planets and their natural satellites (such as our moon); dwarf planets; asteroids and comets. Our solar system is located in an outward spiral of the Milky Way galaxy.

http://solarsystem.nasa.gov/planets/index.cfm

Wayback Machine API

```
curl http://archive.org/wayback/
available?url=solarsystem.nasa.gov/
planets/profile.cfm?
Object=Dwarf&Display=Sats
```

Wayback Machine API


```
{
  archived_snapshots: {
    closest: {
      available: true,
      url: “http://web.archive.org/web/
20150711221849/https://
solarsystem.nasa.gov/planets/profile.cfm?
Object=Dwarf",
      timestamp: "20150711221849",
      status: "200"
    }
  }
}
```

Wayback Machine API

{"archived_snapshots":{}}

404 Handler

Free "404: File Not Found" Handler for Webmasters to Improve User Experience

Posted on October 24, 2013 by internetarchive

The Internet Archive today is launching a free service to help webmasters improve their user experience by augmenting their website's 404 Page Not Found page to link to the Wayback Machine in the case that it has it. Therefore users trying to get to any pages that might have been on a previous version of your website will now be given the option to go to the Wayback Machine.

To embed a link to the Wayback Machine on your site's 404 pages, just include this line in your error page:

```
<div id="wb404"/>
<script src="https://archive.org/web/wb404.js"> </script>
```

If an archived page is not found, then nothing will appear, if it is found, then your user will see:

INTERNET ARCHIVE WaybackMachine

Would you like to see an archived version of this page in the Internet Archive's Wayback Machine?

https://blog.archive.org/2013/10/24/web-archive-404-handler-for-webmasters/

Open Library API

https://openlibrary.org/developers/api

Open Library API

- Query the Open Library database
- View record information
- Edit record information
- View record history

Open Library API

```
http://openlibrary.org/subjects/pluto.json
?limit=1
```

Open Library API


```
{
  works: [
  {
    printdisabled: false,
    cover_id: null,
    ia_collection: [ ],
    has_fulltext: false,
    edition_count: 2,
    checked_out: false,
    title: "The planets",
    public_scan: false,
    lendinglibrary: false,
    lending_edition: "",
    overdrive: "",
    first_publish_year: null,
    key: "/works/OL2715443W",
authors: [
    {
      name: "Heather Couper",
      key: "/authors/
OL397296A"
    },
    {
      name: "Nigel Henbest",
      key: "/authors/
OL449922A"
    }
    ],
    ia: null,
    lending_identifier: "",
    subject: [
      "Solar system",
      "Juvenile literature",
      "Planets"
    ]
  }
  ],
  subject_type: "subject",
  work_count: 13,
  key: "/subjects/planets",
  name: "planets"
}
```

Evergreen

http://evergreen-ils.org/

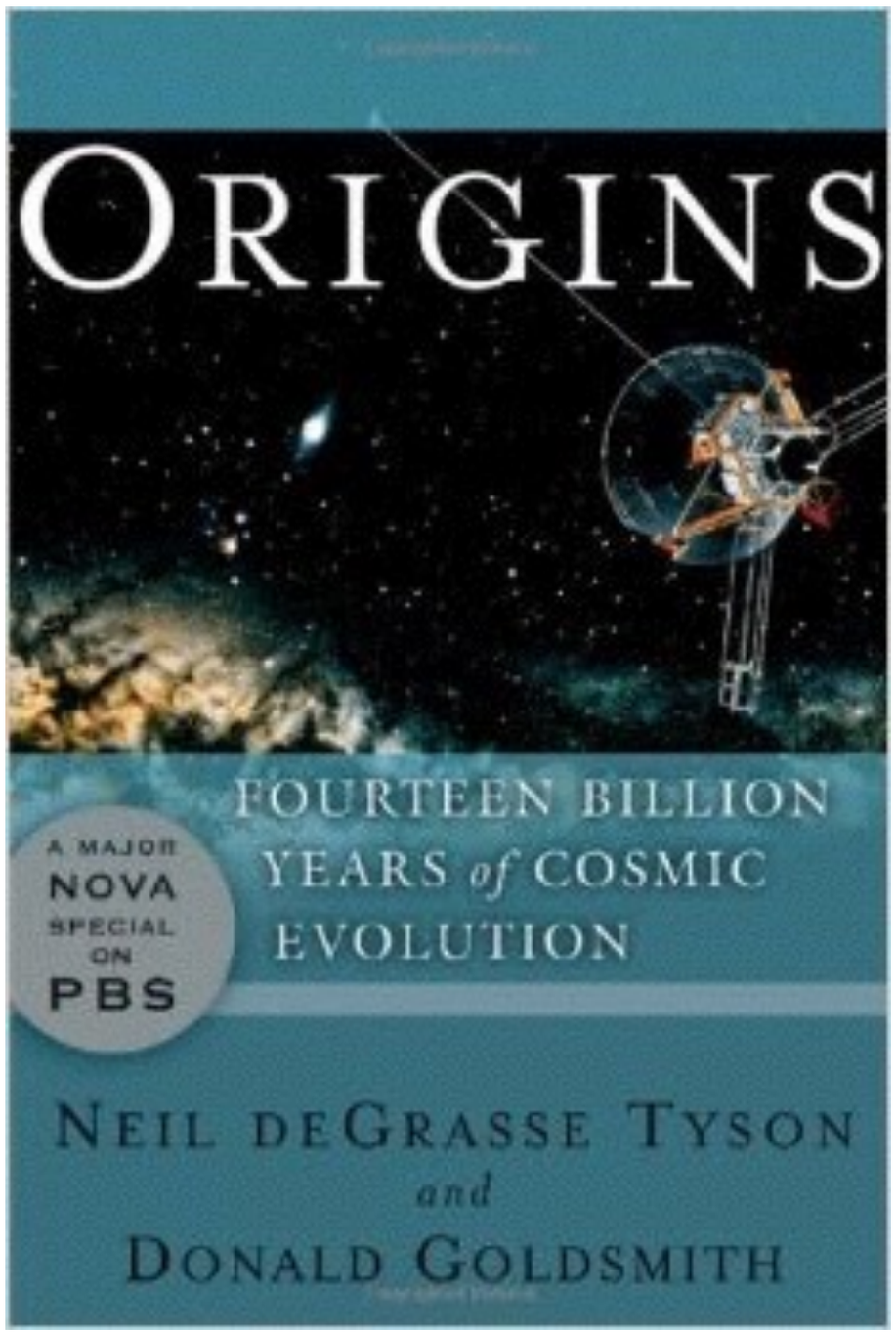
Do-We-Want-It? API

http://want.archive.org/

Do-We-Want-It? API

```
http://want.archive.org/api?isbn=<isbn10 or isbn13>
```

Do-We-Want-It? API



http://want.archive.org/api?isbn=978-0393350395

Do-We-Want-It? API

```
{
  status: "success",
  result: "1",
  description: "want_for_ia_pa",
  identifier: "-1"
}
```

Do-We-Want-It? API

```
Response keys:
  status
    fail    - We failed to process the request. The submitted ISBN was
              invalid.
    success - The request was serviced successfully.
  result
    -1 -  We failed to process the request. The submitted ISBN was
          invalid.
    0  -  The request was serviced successfully, but we have two copies,
          and do not want more.
    1  -  The request was serviced successfully, and we have no copies.
          We want it.
    2  -  The request was serviced successfully, and we have 1 copy
          already. We want it.

  description - "human-parsable" description, with respect to the above:
    failure
    do_not_want
    want_for_ia_pa
    want_for_alt_pa

  identifier - String when result = 2. This is the identifier already
  assigned on the cluster for that ISBN
    -1 - No identifier designated for this ISBN.
    <string> - Designated identifier for this ISBN.
```

IA Search API

https://archive.org/advancedsearch.php

IA Search API

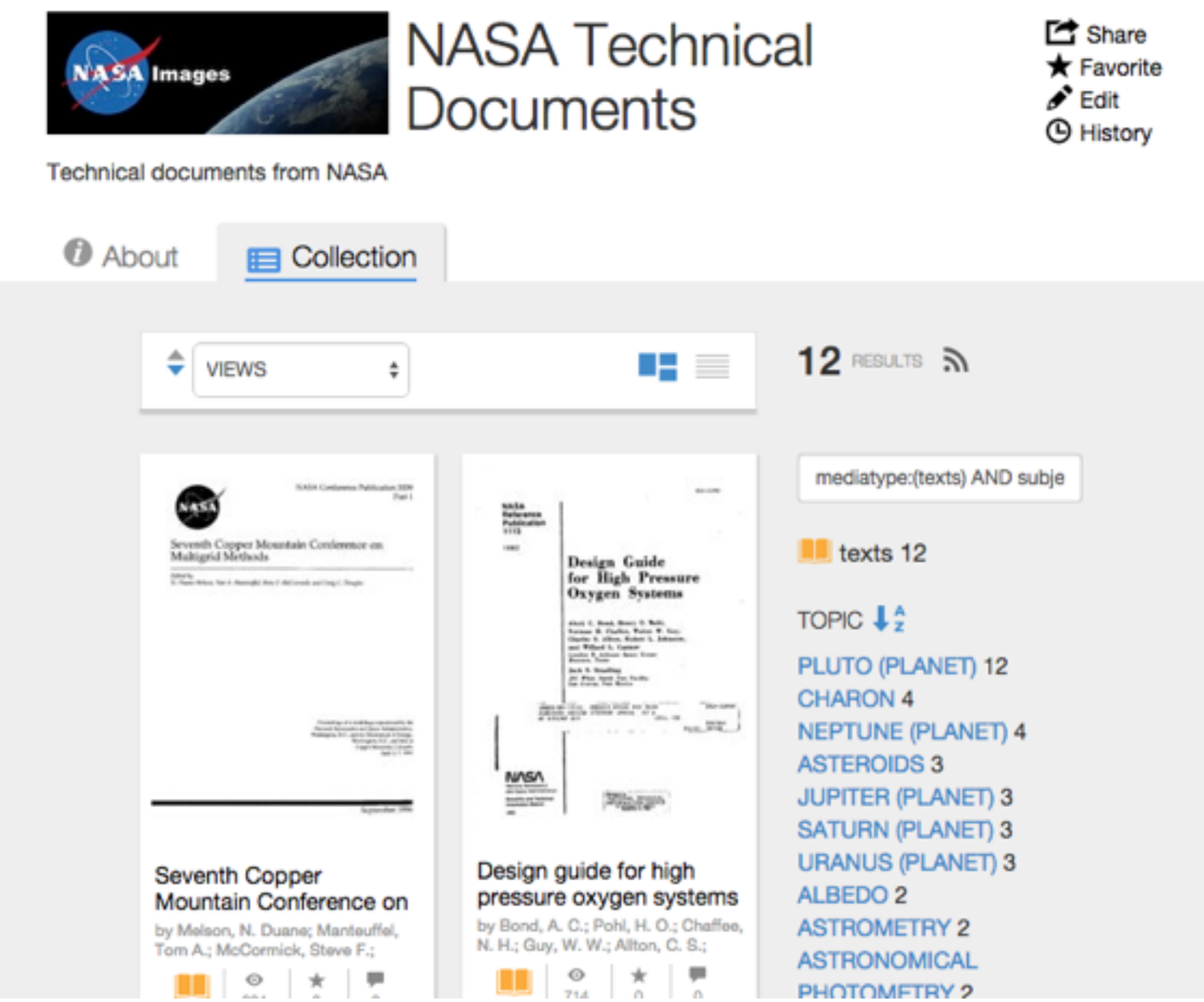
Advanced Search

This form allows you to perform an advanced search. You only need to fill in one field below. This can be any field. If you select "not" as your match criteria, you must select one other field.

	Any field:	contains	
AND	Title:	contains	
AND	Creator:	contains	
AND	Description:	contains	
AND	Collection:	is	All collections
AND	Mediatype:	is	texts
AND	subject	contains	pluto
AND	Custom field	contains	
AND	Custom field	contains	
AND	Date:	YYYY MM DD	
AND	Date range:	YYYY MM DD TO YYYY MM DD	

Search

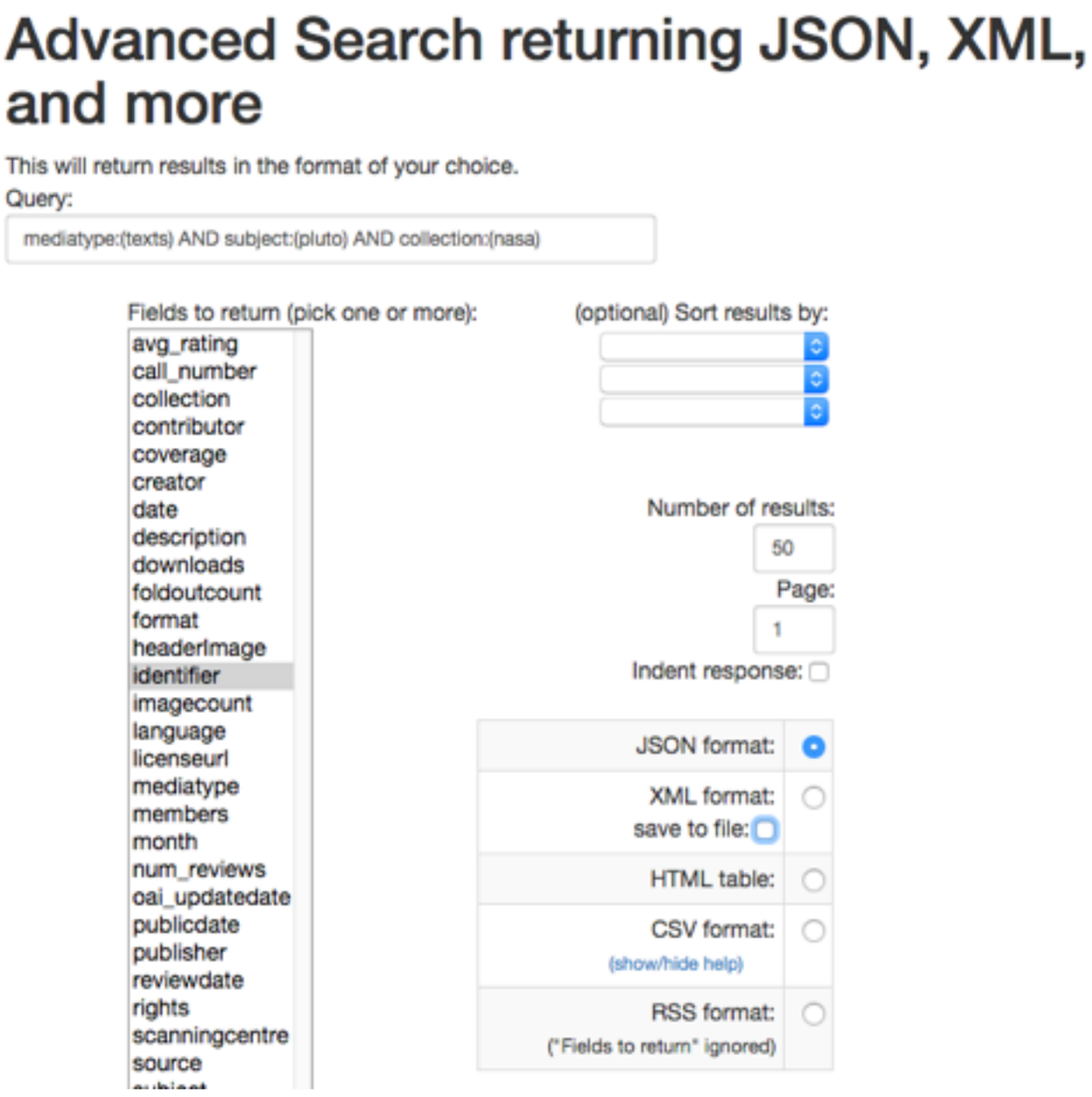
IA Search API



The search terms:
mediatype:(texts) AND subject:(pluto)

https://archive.org/search.php?query=mediatype%3A%28texts%29%20AND%20subject%3A%28pluto%29

IA JSON API



https://archive.org/advancedsearch.php?q=mediatype%3A%28texts%29+AND+subject%3A%28pluto%29+AND+collection%3A%28nasa%29&fl%5B%5D=identifier&sort%5B%5D=&sort%5B%5D=&sort%5B%5D=&rows=50&page=1&output=json&callback=callback

IA JSON API

```
callback(
  {
    responseHeader: {
      status: 0,
      QTime: 108,
      params: {
        json.wrf: "callback",
        wt: "json",
        rows: "50",
        qin: "mediatype:(texts) AND subject:(pluto)",
        fl: "identifier",
        start: "0",
        q: "mediatype:(texts ) AND subject:(pluto )"
      }
    },
    response: {
      numFound: 32,
      start: 0,
      docs: [
        ...
        {
          identifier: "nasa_techdoc_20050060913"
        },
        ...
      ]
    }
  }
)
```

IA JSON API

- Lucene-based
  - Grouping
  - Fuzzy queries
  - Relevance boost
  - Date ranges
  - Etc.

IA Metadata API

http://blog.archive.org/2013/07/04/metadata-api/

IA Metadata API

http://archive.org/metadata/nasa_techdoc_20050060913/metadata

IA Metadata API


```
{
  result: {
    identifier: "nasa_techdoc_20050060913",
    date: "2004",
    description: "Terra MODIS 250 m observations are being applied to a Suspended Sediment Concentrat
algorithm that is under development for coastal case 2 waters where reflectance is dominated by sedim
entrained in major fluvial outflows. An atmospheric correction based on MODIS observations in the 500
resolution 1.6 and 2.1 micron bands is used to isolate the remote sensing reflectance in the MODIS 25
resolution 650 and 865 nanometer bands. SSC estimates from remote sensing reflectance are based on ac
inherent optical properties of sediment types known to be prevalent in the U.S. Gulf of Mexico coasta
present our findings for the Atchafalaya Bay region of the Louisiana Coast, in the form of processed
over the annual cycle. We also apply our algorithm to selected sites worldwide with a goal of extendi
utility of our approach to the global direct broadcast community.",
    document-source: "CASI",
    documentid: "20050060913",
    nasa-center: "Goddard Space Flight Center",
    online-source: "http://wayback.archive-it.org/1792/20100127084754/http://hdl.handle.net/2060/2005
    original-nasa-rights: "Unclassified; No Copyright; Unlimited; Publicly available; Progress Report
    title: "Estimating Coastal Turbidity using MODIS 250 m Band Observations",
    updated-added-to-ntrs: "2008-06-02",
    year: "2004",
    collection: "nasa_techdocs",
    contributor: "NASA",
    language: "eng",
    licenseurl: "http://creativecommons.org/licenses/publicdomain/",
    mediatype: "texts",
    rights: "Public Domain",
…
```

IA Metadata API

It can write metadata, too!

But Wait!

There’s more!

IAS3 API

https://github.com/vmbrasseur/IAS3API

IAS3 API

Create items in Internet Archive, upload files to those items, maintain the metadata for the items, and download from any publicly-available Internet Archive items.

IAS3 API

- Drop-in replacement for the Amazon S3 API
- Pick your favorite S3 library/client, change the server to s3.us.archive.org, and you’re good to go.

IAS3 API

```
curl --location --header 'x-amz-auto-make-bucket:1' \
    --header 'x-archive-meta01-collection:nasa_techdocs' \
    --header 'x-archive-meta-mediatype:movie' \
    --header \
    'x-archive-meta-title:Pluto Fly By' \
    --header "authorization: LOW $accesskey:$secret" \
    --upload-file new-horizon.mp4 \
http://s3.us.archive.org/pluto-new-horizon/new-horizon.mp4
```

RECAP & Global Public Safety Codes

https://archive.org/details/usfederalcourts

https://archive.org/details/publicsafetycode

https://archive.org/details/nasa

Life is short.
What if I don’t want to learn S3?

ia-wrapper

https://github.com/jjjake/ia-wrapper

iadownload

```
#! /usr/bin/env python

# iadownload: Download all files in a collection or item
# Copyright 2014 VM Brasseur

import os
import sys
import internetarchive
import pprint
import argparse
import json
```

https://github.com/vmbrasseur/iadownload

iaupload

README.md

iaupload

NAME

`iaupload` - Upload files to Internet Archive.

SYNOPSIS

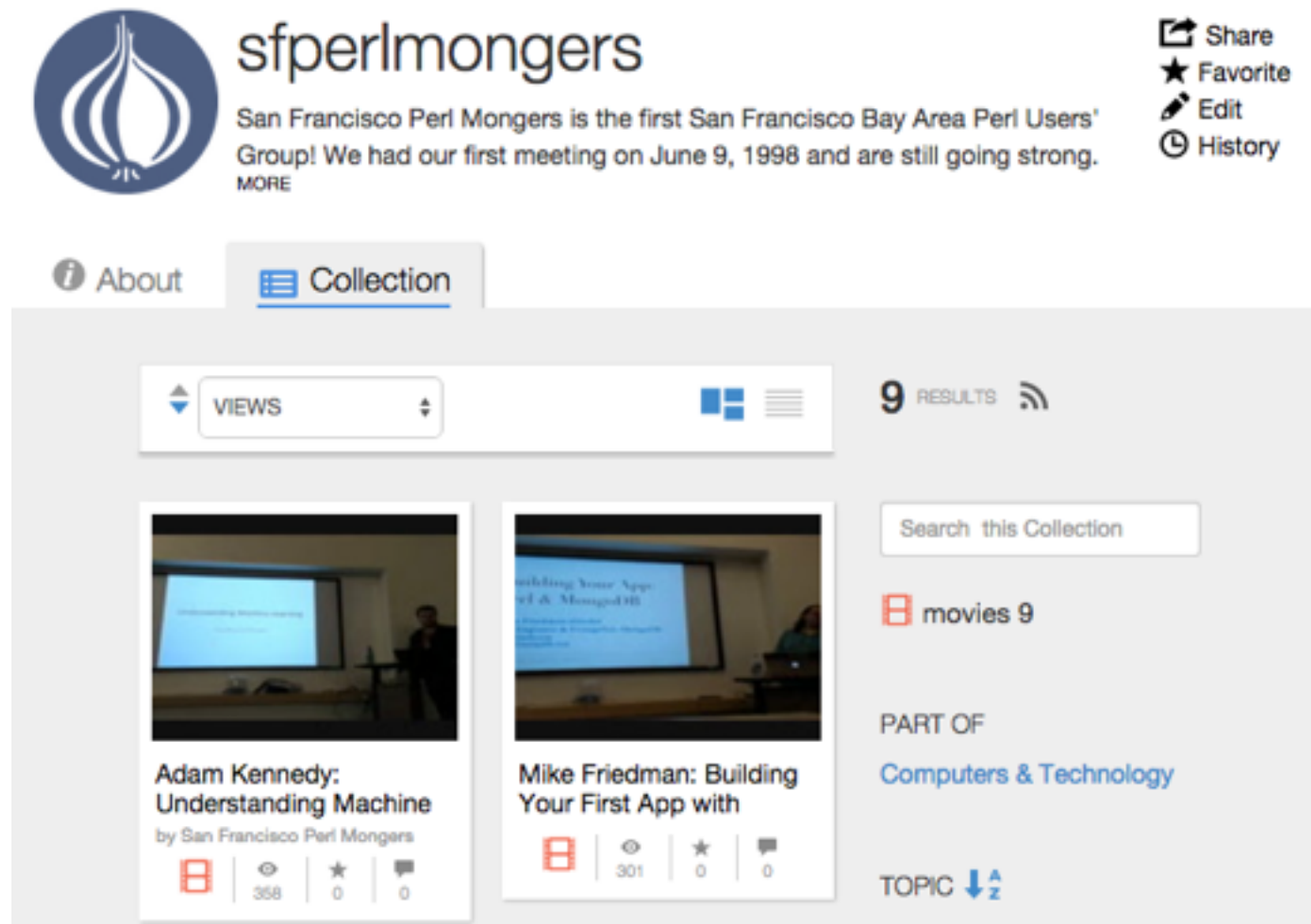
```
usage: iaupload.py [-h] [--config CONFIG] [--metadata METADATA]
       --identifier IDENTIFIER file [file ...]
```

DESCRIPTION

`iaupload` is a Python script I'm using to upload files to the San Francisco Perl Mongers Collection on Internet Archive.

This script uses the internetarchive Python Library.

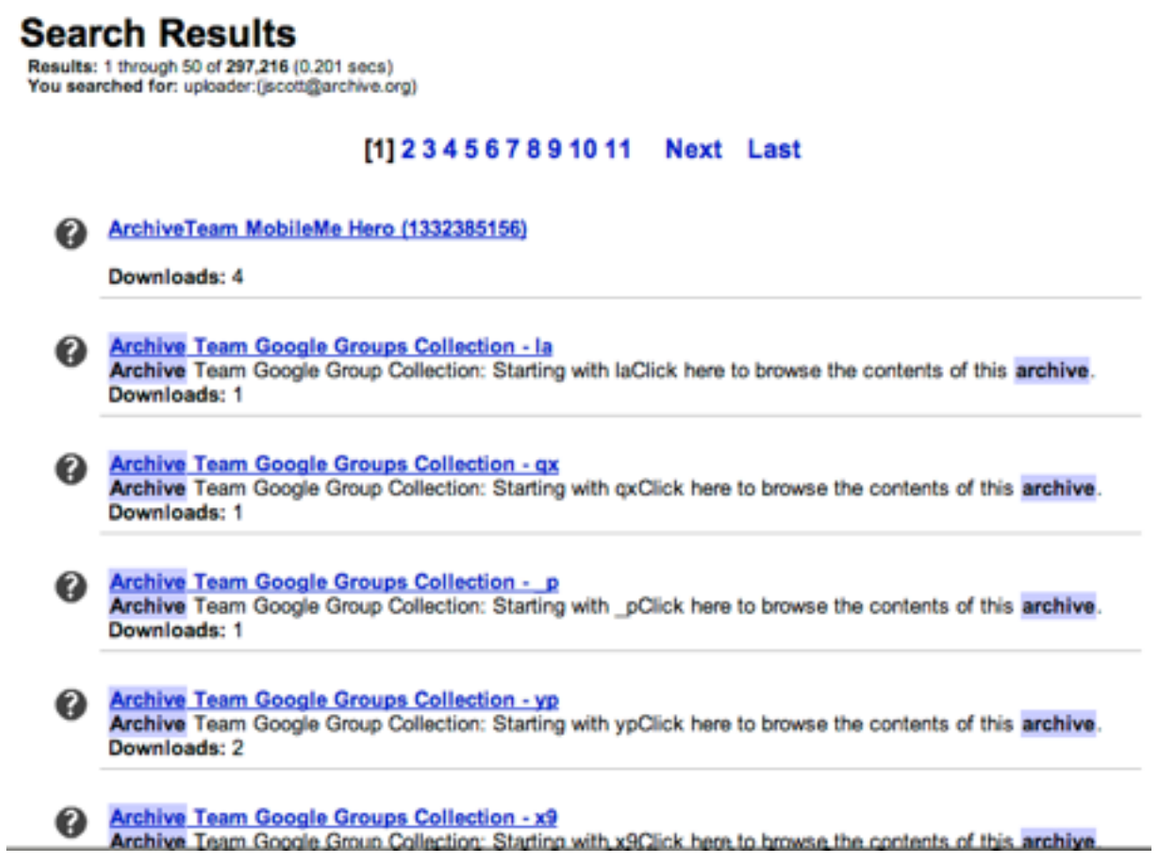
OPTIONS



https://github.com/vmbrasseur/iaupload

http://archive.org/details/sfperlmongers

Saving All The Things



Search Results

Results: 1 through 50 of **297,216** (0.201 secs)
You searched for: uploader:(jscott@archive.org)

[1] 2 3 4 5 6 7 8 9 10 11 Next Last

ArchiveTeam MobileMe Hero (1332385156)
Downloads: 4

Archive Team Google Groups Collection - la
Archive Team Google Group Collection: Starting with laClick here to browse the contents of this **archive**.
Downloads: 1

Archive Team Google Groups Collection - qx
Archive Team Google Group Collection: Starting with qxClick here to browse the contents of this **archive**.
Downloads: 1

Archive Team Google Groups Collection - _p
Archive Team Google Group Collection: Starting with _pClick here to browse the contents of this **archive**.
Downloads: 1

Archive Team Google Groups Collection - yp
Archive Team Google Group Collection: Starting with ypClick here to browse the contents of this **archive**.
Downloads: 2

Archive Team Google Groups Collection - x9
Archive Team Google Group Collection: Starting with x9Click here to browse the contents of this **archive**.

Source: Mirka23 on Flickr

Source: Wikimedia Commons

Recap

- Wayback API
- Open Library APIs
- Do-We-Want-It? API
- Search/JSON APIs
- Metadata API
- IAS3 API
- ia-wrapper

Nikon
D700
FX

Those Links Again…

- https://github.com/internetarchive
- https://archive.org/details/lca2016-ia-apis
- https://www.zotero.org/groups/internet_archive_-_open_apis_and_examples/items

Donate to Internet Archive!

http://archive.org/donate/

THANK YOU!

Internet Archive is HIRING! (SF or Remote)

- *Senior Wayback Machine Engineer*
- *Senior Dev Ops Engineer*
- *Senior Cluster Storage & Computing Engineer*
- *Senior Python Engineer*
- *Web Application Developer*

We need programmers to help us change (and preserve) the world!

VISIT US
Free lunch Fridays at noon
300 Funston Ave
San Francisco

alexis@archive.org